Ethnicity specific microbial signatures in the oral microbiome

Matthew R. Mason

I. Introduction

Personalized medicine is based on the paradigm that definitions of health and disease vary significantly among individuals. Since the ultimate goal of therapy is the attainment of health, delineating health in different groups and cohorts is essential to develop personalized therapeutics. Two of the most common diseases to affect humans, namely caries and periodontal diseases, result from bacterial infections in the oral cavity. Additionally, evidence is emerging to show that oral microbial communities play a critical role in the pathogenesis of oral cancer(Tateda, Shiga et al. 2000; Tezal, Sullivan et al. 2009). In order to understand the role bacteria play in increasing susceptibility to these diseases, it is important to examine the factors that contribute to microbial colonization in health. Within minutes after birth, bacteria colonize the oral cavity and form stable microbial communities in several niches within this ecosystem(Socransky and Manganiello 1971). It is known that the host genotype plays an important role in influencing microbial colonization(Stewart, Chadwick et al. 2005). Since ethnicity is a fundamental component of host genotype, we investigated if ethnicity is a determinant of oral bacterial colonization.

II. Materials and Methods

A. Study population

Approval for this study was obtained from the Office of Responsible Research Practices at The Ohio State University (2008H0122). Periodontally healthy individuals over 18 years of age were recruited from those responding to recruiting campaigns. All subjects interested in the study were emailed a screening questionnaire. This electronic interview served to exclude subjects who were below 18 years of age and satisfy the exclusion criteria listed. Subjects who reported diabetes, HIV, pregnancy, immunosuppressant medications, bisphosphonates or

steroids, current smoking history, current orthodontic therapy, antibiotic therapy or professional cleaning within the previous 3 months, as well as those who required antibiotic coverage before dental treatment, and those who did not meet the ethnicity requirements were excluded from this study. A total of 192 subjects successfully completed the study. Each ethnic group, including African American, Caucasian, Chinese, and Latino, was represented by 48 subjects.

B. Initial clinical screening

Qualifying subjects participated in a periodontal examination to ensure that they satisfied the clinical criteria for inclusion into the study. All subjects were examined by calibrated periodontists. Gingival and plaque indices were recorded throughout the mouth using a PCP-UNC 15 probe. Subjects with at least 20 natural non-carious teeth, ≤3 mm probing pocket depths at all sites (indicative of healthy gums), average pre-brushing plaque score of ≥1.9 (Quigley-Hein modification of the Turesky Plaque Index TPI)(Turesky, Gilmore et al. 1970) and a Loe and Silness gingival index (GI)(Loe and Silness 1963) of ≤1 were selected using this clinical examination.

C. Informed consent and inclusion into study

Each subject who qualified for the study was explained the purpose and procedures of the research. They were given an option to exit the research at this point. If this option was chosen, all data collected during initial screening was destroyed. Informed consent and HIPPA regulations were also explained.

D. Sample Collection

Saliva was collected by expectorating into a sterile 1.5 mL tube using a methodology as previously described(Navazesh 1993). Briefly, subjects will be asked to collect saliva in their mouth for 3 minutes and then continuously drool into a tube for 3 minutes. This method will

allow us to collect unstimulated saliva that will contain significantly greater numbers of bacteria than simply spitting into a tube. Supragingival plaque was collected from interproximal sites using scalers. Following supragingival plaque removal, the area was isolated and subgingival plaque was collected by inserting endodontic paperpoints (Caulk Dentsply) into the interproximal gingival sulci of 10 randomly selected teeth. All the paper point and scaler samples were pooled.

E. DNA isolation

A previously described methodology for DNA isolation was used(McClellan, Griffen et al. 1996). For saliva samples, 50μl of saliva was added to 200μl of phosphate buffered saline (PBS) before preceding with isolation using a Qiagen MiniAmp kit (Valencia, CA) according to manufacturer's instructions. For plaque samples, bacteria were removed from the paper points by adding 200 μl of phosphate buffered saline (PBS) and vortexing for 1 minute. The paper points were then removed, and DNA isolated using a Qiagen MiniAmp kit (Valencia, CA) according to the manufacturer's instructions.

F. t-RFLP analysis

Bacterial 16S rRNA genes were amplified using 22 cycles of PCR with fluorescent-labeled broad range bacterial primers A18-FAM (5'- TT TGA TCC TGG CTC AG–FAM-3') and 317-HEX (5'- FAM-AAG GAG GTG ATC CAG GC -3') (Applied Biosystems, Foster City, CA). The cycling conditions have previously been described(Kumar, Griffen et al. 2005). The amplicons were purified using a Qiaquick kit (Qiagen, Valencia, CA). Restriction digestion was carried out with 10μl of standardized, purified PCR product and 10 U of *Msp I* in a total volume of 20μl at 37°C for three hours. 10μl of the digestion product was purified using AMPure beads (Agencourt Bioscience Corporation, Beverly, MA) according to the manufacturer's protocol and

eluted in 50µl water. 5µl of the purified product was denatured with 10µl of deionized formamide and mixed with 0.2µl GeneScan 1200 LIZ size standard (Applied Biosystems, Foster City, CA). Fragment lengths were determined on an AB 3730 DNA Analyzer in GeneScan mode. The number of peaks as well as the height and area of each peak; reflecting the sizes and intensities of the terminal fragments were determined using the GeneMapper 4.0 Software. Peak areas were standardized by converting the raw values to a proportion of the total area as previously described(Rees, Baldwin et al. 2004). Peaks representing less than 1% of the total area were assigned a value of zero and the percentages of the remaining peaks recalculated. A variance stabilizing transformation was used to create normal distribution of the data(Shchipkova, Nagaraja et al. 2010). The proportion (p) of each peak in the community of each subject was expressed as $X = \sin^{-1}(\sqrt{p})$ and were used for nonmetric multidimensional scaling (NMDS) computed within SPSS (IBM, Armonk, NY). Visualization was carried out with JMP (SAS Institute Inc., Cary, NC).

G. Pyrosequencing

Multiplexed bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP) was performed using the Titanium platform (Roche Applied Science, Indianapolis, IN) as previously described(Dowd, Sun et al. 2008) in a commercial facility (Research and Testing Laboratories, Lubbock, TX). Briefly, a single step PCR with broad-range universal primers and 22 cycles of amplification was used to amplify the 16S rRNA genes as well as to introduce adaptor sequences and sample-specific bar-code oligonucleotide tags into the DNA. Two regions of the 16S rRNA genes were sequenced: V1-V3 and V7-V9. The primers used for sequencing have been previously described (Kumar et al, PlosONE). Adaptor sequences were trimmed from raw data with 98% or more of bases demonstrating a quality control of 30 and sequences binned into

individual sample collections based on bar-code sequence tags, which were then trimmed. The resulting files were denoised with Pyronoise(Quince, Lanzen et al. 2011) and depleted of chimeras using B2C2 (http://www.researchandtesting.com/B2C2.html). Sequences <300bp were discarded and the rest were clustered into species-level operational taxonomic units (s-OTUs) at 96% sequence similarity and assigned a taxonomic identity by alignment to locally hosted version of the Greengenes database(DeSantis, Hugenholtz et al. 2006) using the Blastn algorithm. Phylogenetic trees were generated by MacVector and visualized using iTOL(Letunic and Bork 2007). Community diversity metrics were computed as previously described(Lozupone, Lladser et al. 2007).

H. Statistical analysis

Shannon diversity index was computed using s-OTU data(Shannon 1997). A variance stabilizing transformation was used to create normal distribution of the data(Shchipkova, Nagaraja et al. 2010). The proportion (p) of each s-OTU in the community of each subject was expressed as $X = \sin^{-1}(\sqrt{p})$ and ANOVA and 2-sample t-tests were used to compare the means of this transformed variable X across groups. Species and genera shared by ethnic groups were identified used to compute both the core microbiome as well as ethnicity-specific microbiomes. Species present in >80% of each ethnic group were considered for analysis. Discriminant analysis of each individual's microbial community was performed using a trained random forest machine learning algorithm carried out with Statistica (StatSoft Inc., Tulsa, OK). To predict the likelihood that an individual was of a certain ethnicity given their microbial signature we calculated the number of subjects in an ethnic group that contained >80% of their respective ethnicity-specific microbiome species divided by the total number of subjects from different ethnicities who also contained >80% of the numerator's ethnicity-specific microbiome species.

Statistical analysis was carried out with JMP (SAS Institute Inc., Cary, NC) and graphics created using R (http://www.r-project.org/).

III. Results

We compared the oral microbial communities of 192 people belonging to four ethnicites: non-Hispanic blacks (AA), non-Hispanic whites (CA), Chinese (CH), and Latinos (LA). These ethnicites were selected since they represent four major races/ethnic groups residing in the United States. All subjects reported both parents and both sets of grandparents to be of the same ethnicity; Chinese and Latino subjects were either immigrants from China and Taiwan, or Central America and Puerto Rico respectively, or first generation residents. All subjects were free of systemic diseases, active caries, and periodontal diseases.

We used terminal restriction fragment length polymorphism (t-RFLP), to compare the 'fingerprints' of the salivary, supragingival and subgingival microbiomes between the four ethnic groups. These environments represent three distinct microbial niches within the oral ecosystem. Supragingival plaque forms on a non-shedding surface that is exposed to mechanical and frictional forces and hence, represents a biofilm where the effects of the environment supercede the effects of the host genotype. The subgingival biofilm on the other hand represents a community that is influenced to a large extent by genetically controlled host-associated factors (for example tooth morphology, epithelial barrier function, and innate and adaptive immune responses). Saliva represents a fluid environment in communication with all oral habitats that shares microbial "fingerprints" with both supragingival and subgingival ecosystems. 16S rRNA genes were digested using restriction enzymes, generating terminal fragments of varying lengths based on sequence variations among the different bacterial species. Thus, the total number of peaks represents the number of unique species present in the community and the area of each

peak represents the abundance of each species. Non-metric Multi-Dimensional Scaling (PROXSCAL NMDS) of the Bray Curtis Similarity Index(Bray and Curtis 1957) was used to examine the association between ethnicity and the microbial composition of the three oral niches. We found significant clustering by ethnicity within the subgingival microbial community, but not the salivary or supragingival communities (Figure 1a-c). These results provided us with early evidence that host genotype plays a vital role in determining the composition of microbial communities. Further studies are warranted to examine the role of host genotype in the establishment of microbial communities in other sheltered niches within the human host.

Based on the clustering, we characterized the bacterial lineages in the subgingival microbiome of 100 individuals using multiplexed 16S pyrotag sequencing. For each sample, variable regions V1-V3 and V7-V9 of the bacterial 16S ribosomal RNA (rRNA) gene were sequenced and combined to create a composite dataset. A total of 633, 601 high-quality, chimera-depleted, classifiable sequences were obtained. These sequences represented 398 species-level operational taxonomic units (s-OTUs) with an average of 149±34 s-OTUs detected in each individual. S-OTU data was used to compute Shannon Diversity and Equitability indices. The Shannon index incorporates both the number of s-OTUs (richness) and relative abundance of each s-OTU (evenness) into a single value. While a Diversity Index of zero represents a mono-species community, a higher value may result either from the presence of several species or from equitable distribution of a few species. Thus, the Equitability index serves to characterize the relative contributions of species richness and evenness to the Diversity index. African Americans had lower Diversity($p=.0006$, ANOVA) (Figure 2a) and Equitability ($p=.0002$, ANOVA) (Figure 2b) indices when compared to the other three ethnic groups. This finding

indicates that African Americans have fewer types of subgingival species than the other ethnicities and that these species are not equally abundant in the community.

The Human Microbiome Project has highlighted the importance of identifying 'core microbiomes' that are common to all healthy individuals in order to understand susceptibility to disease. We found eight s-OTUs (2%) that were present in all 100 individuals (Figure 3a). Moreover, 8% of the 398 s-OTUs were detected in 90% of individuals and over a third of the s-OTUs were shared by half of the subjects (Figure 3a). These findings support the existence of a 'core microbiome' within the subgingival habitat. However, we also found the existence of s-OTUs unique to each ethnicity (Figure 3b) indicating a possible ethnicity-based selection in the composition of the subgingival microbial community. Furthermore, half of the eight s-OTUs present in all subjects showed significant differences in abundances between ethnicities (Figure 3c) lending further support to the fact that ethnicitiy plays a role in determining the composition of the subgingival microbiome.

Analysis of the datasets at the genus level further served to confirm this finding, since 33 of the 77 genera demonstrated significant differences in abundance between the ethnic groups ($p<0.05$, ANOVA) (Figure 4). This suggests that distinct bacterial lineages contribute to the composition of the subgingival communities in different ethnicities. In addition, we found that the subgingival microbial fingerprint is successfully able to discriminate between the four ethnicities using a Random Forest machine-learning classifier. The Random Forest classifier uses a training dataset to develop an educated classification algorithm, which is then applied to a test dataset to examine the accuracy, sensitivity and specificity of the prediction. We found that taken as a whole, the subgingival microbial community was able to predict an individual's ethnicity with a 62% accuracy, 58% sensitivity and 86% specificity (Figure 5). The classifier was able to

predict African Americans with a 100% sensitivity and 74% specificity and Caucasians with a 50% sensitivity and 91% specificity (Figure 5). This is interesting because although African Americans and Caucasians shared similar environmental factors including food, nutrition, and lifestyle over several generations, they demonstrated distinct microbial communities. This suggests that the host genotype influences the subgingival microbial community to a greater extent than shared environment; "nature" appears to win over "nurture" in shaping this community.

We then investigated if the presence a consortium of selected microbial species could be surrogates of the total microbiome to predict an individual's ethnicity. To do this, we identified species that were present in at least 80% of the subjects within each ethnicity (Figure 6). Next, we estimated the likelihood that the ethnicity-specific microbial consortia will predict an individual's ethnicity. This method demonstrated a prediction likelihood of 65% for African Americans, 45% for Caucasians, 33% for Chinese, and 47% for Latinos (Table 1). In light of the fact that several oral diseases including periodontitis are more prevalent among African Americans when compared to Caucasians, these findings suggest that bacterial colonization in health may be an indicator of susceptibility to future disease.

IV. Discussion

Subgingival biofilms are a major etiologic agent for periodontal diseases. The progression from health to disease is initiated by a shift in the composition of the biofilm towards increasing ratios of certain species and the subsequent host response to this changing community(Matthews, Joshi et al. 2012). Thus, the host response defines the pathogenicity of a species. As with other ecosystems, in a state of health certain bacteria with low immunogenic potential (early colonizers) colonize the subgingival crevice in large numbers limiting the

abundance of immunogenic species, a phenomenon known as niche saturation. When the ratio of high-immunogenic microbes begins to increase and compete out the health-compatible community, the host-mediated immune response is amplified and results in disease. The data presented here suggest that the health-compatible bacteria exist in signifcantly different ratios among different ethnicities. This is especially evident in African Americans and may contribute to the increased susceptibility to disease that has been observed in this cohort(Albandar 2002; Borrell, Burt et al. 2005).

In summary, the work presented here demonstrates the existence of ethnicity-specific subgingival microbiomes that are characterized by differing bacterial lineages and varying diversities. It is possible that these health-associated ethnicity-specific microbial communities may predispose individuals to future disease and warrants further examination.

Albandar, J. M. (2002). "Periodontal diseases in North America." Periodontol 2000 **29**: 31-69.
Borrell, L. N., B. A. Burt, et al. (2005). "Prevalence and trends in periodontitis in the USA: the [corrected] NHANES, 1988 to 2000." Journal of dental research **84**(10): 924-930.
Bray, J. R. and J. T. Curtis (1957). "An ordination of upland forest communities of southern Wisconsin." Ecological Monographs **27**: 325-349.
DeSantis, T. Z., P. Hugenholtz, et al. (2006). "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB." Appl Environ Microbiol **72**(7): 5069-5072.
Dowd, S. E., Y. Sun, et al. (2008). "Bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP) for microbiome studies: bacterial diversity in the ileum of newly weaned Salmonella-infected pigs." Foodborne Pathog Dis **5**(4): 459-472.
Kumar, P. S., A. L. Griffen, et al. (2005). "Identification of candidate periodontal pathogens and beneficial species by quantitative 16S clonal analysis." J Clin Microbiol **43**(8): 3944-3955.
Letunic, I. and P. Bork (2007). "Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation." Bioinformatics **23**(1): 127-128.
Loe, H. and J. Silness (1963). "Periodontal Disease in Pregnancy. I. Prevalence and Severity." Acta Odontol Scand **21**: 533-551.
Lozupone, C., M. E. Lladser, et al. (2007). "UniFrac: an effective distance metric for microbial community comparison." Isme J **5**(2): 169-172.
Matthews, C. R., V. Joshi, et al. (2012). "Host-Bacterial Interactions During Induction and Resolution of Experimental Gingivitis in Current Smokers." Journal of periodontology.
McClellan, D. L., A. L. Griffen, et al. (1996). "Age and prevalence of Porphyromonas gingivalis in children." J Clin Microbiol **34**(8): 2017-2019.
Navazesh, M. (1993). "Methods for Collecting Saliva." Annals of the New York Academy of Sciences **694**(1): 72-77.
Quince, C., A. Lanzen, et al. (2011). "Removing noise from pyrosequenced amplicons." BMC Bioinformatics **12**(1): 38.
Rees, G. N., D. S. Baldwin, et al. (2004). "Ordination and significance testing of microbial community composition derived from terminal restriction fragment length polymorphisms: application of multivariate statistics." Antonie Van Leeuwenhoek **86**(4): 339-347.
Shannon, C. E. (1997). "The mathematical theory of communication. 1963." MD Comput **14**(4): 306-317.
Shchipkova, A. Y., H. N. Nagaraja, et al. (2010). "Subgingival Microbial Profiles of Smokers with Periodontitis." J Dent Res.
Socransky, S. S. and S. D. Manganiello (1971). "The oral microbiota of man from birth to senility." J Periodontol **42**(8): 485-496.
Stewart, J. A., V. S. Chadwick, et al. (2005). "Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children." J Med Microbiol **54**(Pt 12): 1239-1242.
Tateda, M., K. Shiga, et al. (2000). "Streptococcus anginosus in head and neck squamous cell carcinoma: implication in carcinogenesis." Int J Mol Med **6**(6): 699-703.
Tezal, M., M. A. Sullivan, et al. (2009). "Chronic periodontitis and the incidence of head and neck squamous cell carcinoma." Cancer Epidemiol Biomarkers Prev **18**(9): 2406-2412.

Turesky, S., N. D. Gilmore, et al. (1970). "Reduced plaque formation by the chloromethyl analogue of victamine C." J Periodontol **41**(1): 41-43.

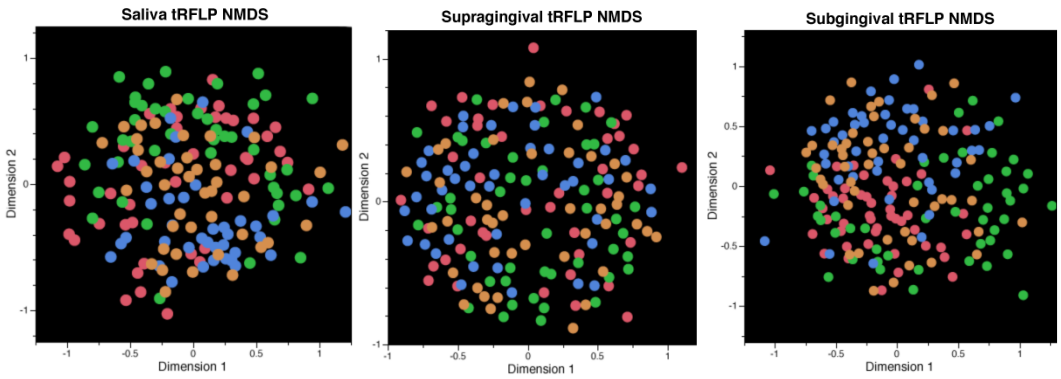

Figure 1. Non-metric multidimensional scaling (NMDS) of tRFLP peak abundance. Saliva is shown in Figure 1A, Supragingival in 1B, and Subgingival in 1C. Significant ethnicity-based clustering was seen in subgingival samples (Subgingival stress value=.09, Saliva stress value=.11, Supragingival stress value=.12). African American samples are indicated by red points, Caucasian samples are indicated by green points, Chinese samples are indicated by blue points, and Latino samples are indicated by orange points.

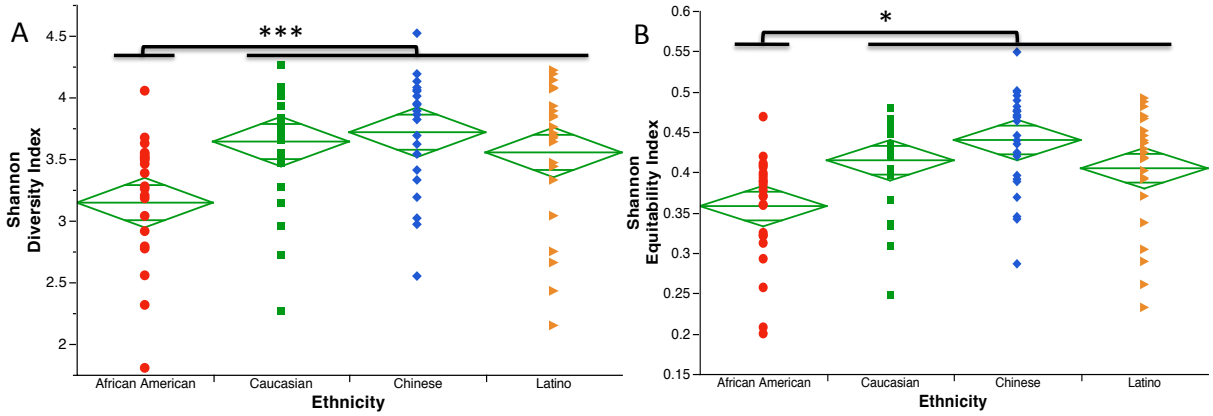

Figure 2. 16s pyrotag sequencing of 100 subjects (n=25 for each ethnicity) reveals reduced diversity (A)(***<.001, ANOVA) and equitability (B)(*<.05, ANOVA) of the subgingival microbial composition in African Americans.

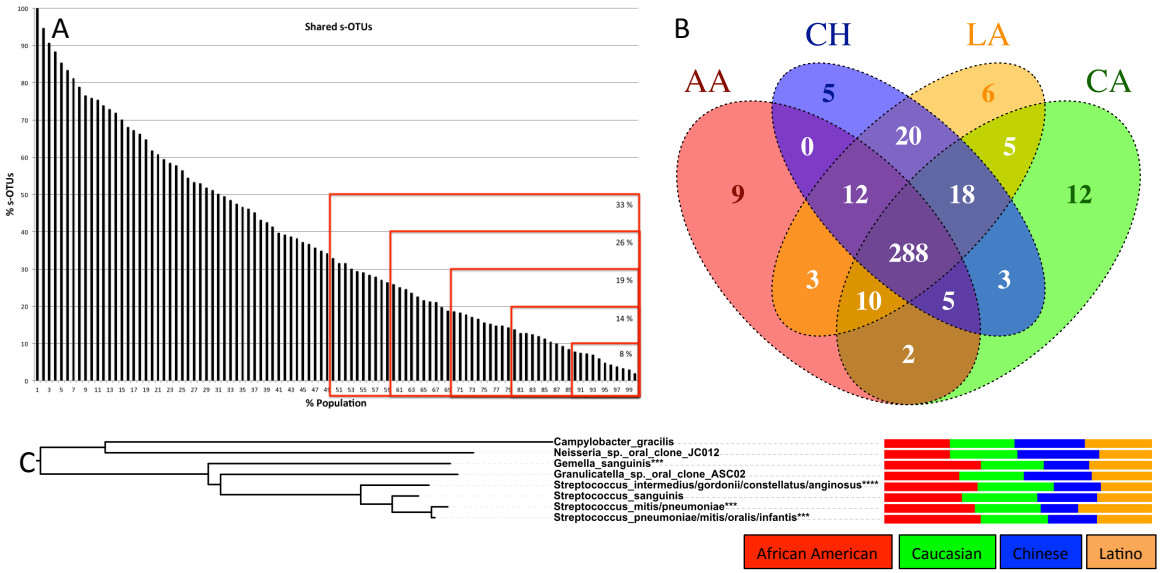

Figure 3.Comparison and identification of shared species and a core microbiome within the study population. Number of species (s-OTUs) shared among the study population (n=100) (A). Venn diagram comparing the number of detected s-OTUs unique to each ethnicity and shared among combinations of various ethnicities (B). Phylogenetic tree of the eight s-OTUs (2%) present in every sample. The connected bars depict mean abundance present in each ethnicity, with four of the s-OTUs present in significantly different abundances (***<.001, ****<.0001) (C).

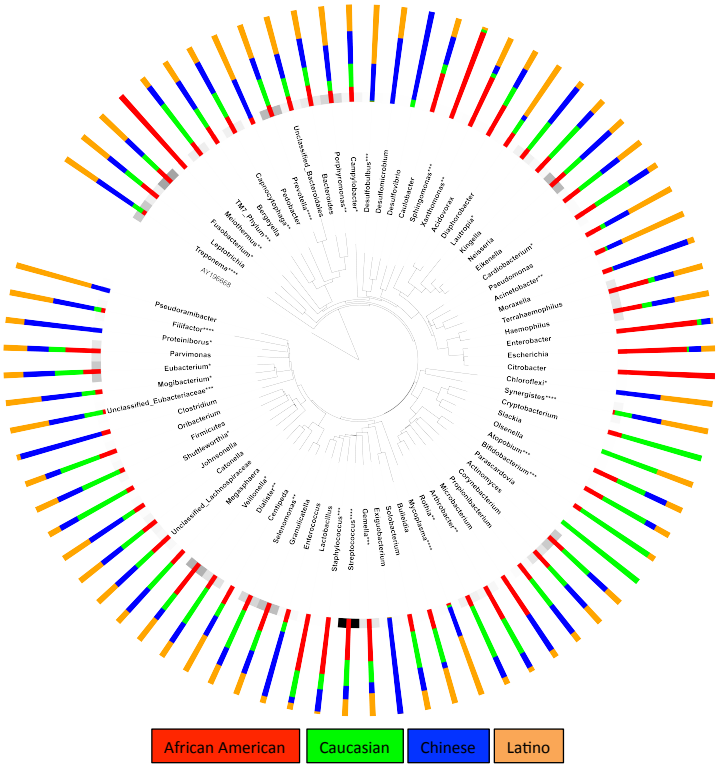

Figure 4. Phylogenetic tree of the 77 genera detected in the study population. A black-white color gradient connected to the genera represents the detection prevalence within the total bacterial population. The colored bars connected to the genera represent the mean abundance within each ethnicity (*<.05, **<.01, ***<.001, ****<.0001)

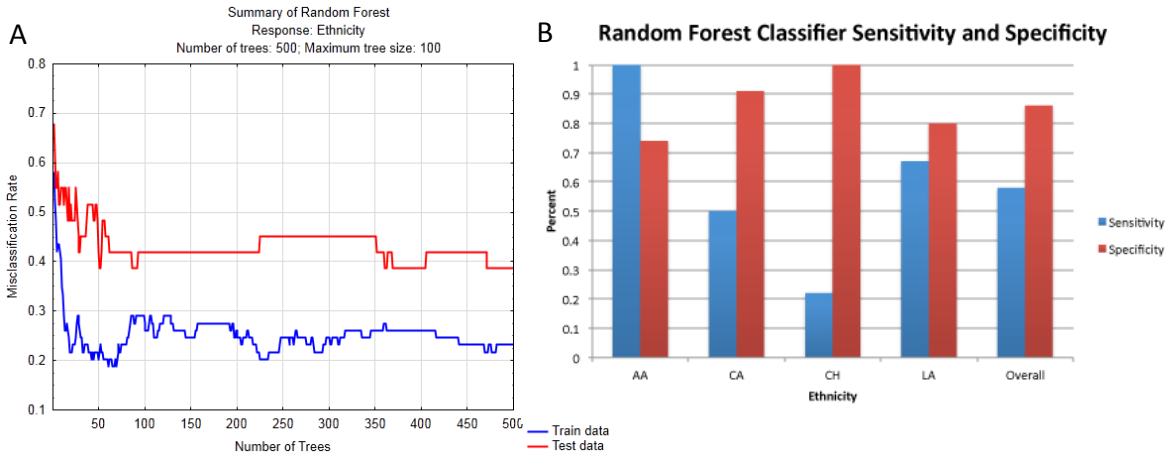

Figure 5. The composition of the subgingival microbial fingerprint successfully discriminates between the four ethnicities. The Random Forest Classifier performance (A) shows the misclassification rate as the number of grown trees increased. The sensitivity and specificity of the random forest classifier for each ethnicity and overall is shown in (B).

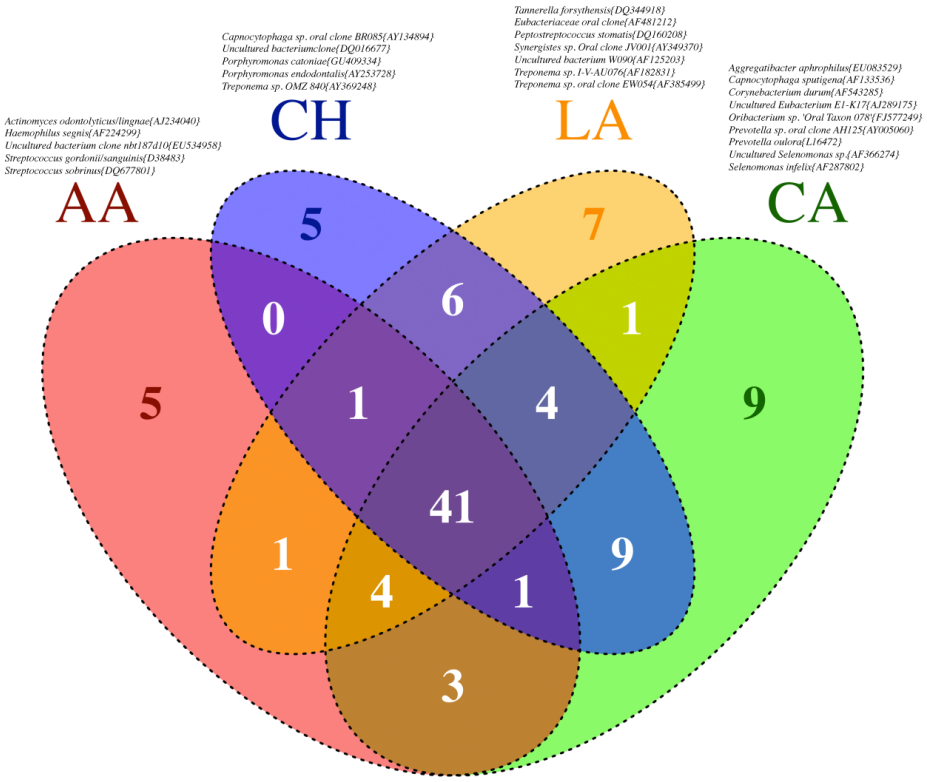

Figure 6. The identification of species present in >80% of subjects revealed an ethnicity specific consortia of species used to estimate the likelihood of an individuals ethnicity in the presence of the consortia.

Ethnicity	# of AA	# of CA	# of CH	# of LA	Likelihood
AA Core Profile	20	13	10	8	65%
CA Core Profile	5	21	12	9	45%
CH Core Profile	9	14	18	14	33%
LA Core Profile	0	3	15	16	47%

Table 1. Using the presence of the ethnicity-specific consortia of s-OTUs (Figure 6), the estimated likelihood of being a certain ethnicity given the presence of an ethnicity-specific consortia of s-OTUs reveals these specific cores are able to successfully predict ethnicity better than chance.